Supreme Court, U.S.
FILED
JAN 19 1993
OFFICE OF THE CLERK

No. 92-102

IN THE
Supreme Court of the United States
OCTOBER TERM, 1992

WILLIAM DAUBERT and JOYCE DAUBERT, Individually and as Guardians Ad Litem for JASON DAUBERT, and ANITA DE YOUNG, Individually and as Guardian Ad Litem for ERIC SCHULLER,
Petitioners,

v.

MERRELL DOW PHARMACEUTICALS, INC.,
Respondent.

On Writ of Certiorari to the United States Court of Appeals for the Ninth Circuit

---

BRIEF AMICUS CURIAE OF PROFESSOR ALVAN R. FEINSTEIN IN SUPPORT OF RESPONDENT

DON M. KENNEDY
LORETTA M. SMITH
RICHARD A. OETHEIMER
Counsel of Record
GOODWIN, PROCTER & HOAR
Exchange Place
Boston, MA 02109
(617) 570-1000
Counsel for Amicus Curiae

January 19, 1993

WILSON - EPES PRINTING CO., INC. - 789-0096 - WASHINGTON, D.C. 20001

BEST AVAILABLE COPY

TABLE OF CONTENTS

	Page
TABLE OF AUTHORITIES	iii
INTEREST OF AMICUS CURIAE	1
SUMMARY OF ARGUMENT	2
ARGUMENT	4
I. JUDICIAL STANDARDS ARE NECESSARY TO ENSURE THE TRUSTWORTHINESS OF SCIENTIFIC EVIDENCE	4
II. PEER REVIEW AND PUBLICATION ESTABLISH NECESSARY STANDARDS AND RIGOROUS DISCIPLINE FOR THE SCIENTIFIC COMMUNITY	7
III. RE-ANALYSIS IS AN ACCEPTABLE SCIENTIFIC METHOD, BUT IT, TOO, SHOULD BE SUBJECTED TO ESTABLISHED SCIENTIFIC STANDARDS AND THE RIGORS OF PEER REVIEW	8
IV. META-ANALYSIS PRESENTS SPECIAL PROBLEMS IN ITS AGGREGATION OF DATA; IT IS VITAL THAT STANDARDS BE ESTABLISHED AND MET BEFORE SUCH DATA ARE PRESENTED TO A JURY	10
V. CONFIDENCE INTERVALS, LIKE STATISTICAL SIGNIFICANCE, ARE A WAY OF MEASURING NUMERICAL STABILITY. EITHER TOOL MUST BE USED WITH ACCEPTED STANDARDS FOR RESULTS TO BE TRUSTWORTHY FOR EITHER SCIENTISTS OR JURIES	12
A. Numerical Stability	12
B. Null Hypothesis and P Values	13
C. Choice of Alpha Boundary	13

TABLE OF CONTENTS—Continued

	Page
D. Role of Sample Size	15
E. Calculation of Confidence Intervals	16
CONCLUSION	20
APPENDIX—IDENTIFICATION OF AMICUS CURIAE	1a

TABLE OF AUTHORITIES

Cases	Page
Daubert v. *Merrell Dow Pharmaceuticals, Inc.*, 951 F.2d 1128 (9th Cir. 1991)	2
Daubert v. *Merrell Dow Pharmaceuticals, Inc.*, 727 F. Supp. 570 (S.D. Cal. 1989)	2
Rules	
Federal Rule of Evidence 702	5
Miscellaneous	
A. Feinstein, R. Horwitz, W. Spitzer, R. Battista, "Coffee and Pancreatic Cancer: The Problems of Etiologic Science and Epidemiologic Case-Control Research," 246 *Journal of the American Medical Association* 957 (1981)	9
R. Fisher, *Statistical Methods For Research Workers* (Edinburgh; Oliver & Boyd, 1925)	14
Z. Gerbarg & R. Horwitz, "Resolving Conflicting Clinical Trials: Guidelines for Meta-Analysis", 41 *Journal of Clinical Epidemiology* 503 (1987)	11
H. Sacks, J. Berrier, D. Reitman, V. Ancona-Berk & T. Chalmers, "Meta-Analyses of Randomized Controlled Trials", 316 *New England Journal of Medicine* 450 (1987)	11
K. Wachtler, "Disturbed By Meta-Analysis", 241 *Science* 1407 (1988)	10
"Whither Meta-Analysis", *The Lancet* (April 18, 1987) 897-898	10

IN THE

Supreme Court of the United States

OCTOBER TERM, 1992

No. 92-102

WILLIAM DAUBERT and JOYCE DAUBERT,
Individually and as Guardians Ad Litem for
JASON DAUBERT, and ANITA DE YOUNG, Individually
and as Guardian Ad Litem for ERIC SCHULLER,
Petitioners,

v.

MERRELL DOW PHARMACEUTICALS, INC.,
Respondent.

**On Writ of Certiorari to the
United States Court of Appeals
for the Ninth Circuit**

BRIEF AMICUS CURIAE OF PROFESSOR ALVAN R. FEINSTEIN IN SUPPORT OF RESPONDENT

INTEREST OF AMICUS CURIAE

Alvan R. Feinstein, M.D. is Sterling Professor of Medicine and Epidemiology at the Yale University School of Medicine, where he is also Director of the Clinical Epidemiology Unit. He is the author of texts entitled *Clinical Judgment*, *Clinical Biostatistics*, *Clinical Epidemiology* and *Clinimetrics*. He is co-editor of the respected *Journal of Clinical Epidemiology*. He has contributed hundreds of articles to the medical and scientific literature over the past 40 years, and helped to pioneer the field of Clinical Epidemiology. A more complete biog-

raphy of the amicus appears in the Appendix to this brief.

Dr. Feinstein submits this brief in support of Respondent Merrell Dow Pharmaceuticals, Inc. to answer the Brief *Amici Curiae* of Professors Kenneth Rothman, Noel Weiss, James Robins, Raymond Neutra and Steven Stellman, in Support of Petitioners (collectively, "Rothman, *et al.*"). For the past quarter-century, Dr. Feinstein has been at the forefront of the field of clinical epidemiology and has advocated developing scientific standards for conducting epidemiologic investigations. Dr. Feinstein disagrees with Rothman, *et al.*'s "free-for-all" approach to the admission of scientific evidence in a court of law and its implicit evisceration of scientific standards. He submits this brief to apprise the Court of the opposing viewpoint, which insists that the standards of scientific and intellectual discipline be recognized and observed, and that scientific standards be maintained and respected when scientific evidence is used in courts of law.[1]

SUMMARY OF ARGUMENT

I have read the Brief *Amici Curiae* of Professors Kenneth Rothman, Noel Weiss, James Robins, Raymond Neutra, and Steven Stellman, in Support of Petitioners ("Rothman Brief"). I have some profound disagreements with its concepts and recommendations. I understand that the issue before this Court concerns the proper role of judges and courts in controlling the flow of scientific evidence to juries in litigation. The approach suggested by Rothman, *et al.* appears to leave no room for the application of suitable scientific standards in evaluating the worth of scientific evidence in litigation. Rothman, *et al.*'s criticisms of the lower courts' decisions[2] entirely miss the point that the matter at issue is not recent scien-

[1] Counsel to the parties have consented to the filing of this brief. Letters of consent have been filed with the Clerk.

[2] *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 727 F. Supp. 570 (S.D. Cal. 1989), *aff'd*, 951 F.2d 1128 (9th Cir. 1991).

tific discovery which can be weighed and evaluated by a jury of scientists, who might well wish to have "all the facts." The information is to be weighed by a jury of lay persons who are subject to understandable confusion by epidemiology. The suggestion that judicial safeguards be abandoned would lead to intellectual anarchy and decision-making by juries based upon suspect or highly unreliable data. Judicial screening of proposed expert evidence is essential to the legal process.

Peer review serves a similar, vital purpose in the scientific community. Although peer review alone does not guarantee the soundness of research, it certainly increases the likelihood that the work is trustworthy because it has been subject to the harsh light of critical appraisal and public exposure. Discarding peer review as a test for scientific reliability would make the judicial process less reliable to the detriment of all. Meta-analysis and re-analysis, techniques urged by Rothman, *et al.*, are not themselves invalid methods. They are, however, subject to abuses, distortions, and inaccuracies, which make their review by peers, prior to unguarded disclosure to a jury, vitally important.

Finally, although it is true that "statistically significant" data are not the only epidemiologic data upon which scientists base inferences, jurors are not themselves practicing scientists. Therefore, there are good reasons for limiting epidemiologic proof in the courtroom to evidence which has met accepted scientific and statistical standards for numerical stability. The tool of "confidence intervals" is not a substitute for the scientific standard of "statistical significance." Confidence intervals can replace P values as the main tool for measuring numerical stability, but they have strengths and weaknesses which are analogous to those of P values. If a specific boundary, such as 90% or 95%, is not set for the confidence levels, the stability of the numbers becomes untrustworthy, and the investigator can arrange the intervals to show almost anything that is desired.

ARGUMENT

I. JUDICIAL STANDARDS ARE NECESSARY TO ENSURE THE TRUSTWORTHINESS OF SCIENTIFIC EVIDENCE.

My first comments refer to the implicit proposal by Rothman, *et al.* that the judge in a legal proceeding act as a mere conduit rather than a careful filter in transferring scientific information to the jury. Rothman, *et al.* say that if they were "on a jury in a civil trial, *amici* would wish to have access to the same sources we consult as practicing scientists. We would want to know all the facts, not just the truncated 'consensus'." Rothman Brief at 13. Thus, Rothman, *et al.* apparently would want anything and everything to be presented directly to the jury, without the judge making any prior appraisal of suitability or pertinence.

I vigorously disagree. First, the phrase "truncated 'consensus' " implies that the published literature is monolithic in viewpoint and grossly inadequate. This implication is false. I do not know of a single disputed issue, particularly in the field of epidemiology, that lacks publications from both sides of the argument (and sometimes from more than just two sides). The total array of publications may be unbalanced in one direction or the other, and the degree of imbalance may be affected by the "consensus" viewpoint. Nevertheless, alternative views are there; they can be found; and they can be introduced into the courtroom. One need not and should not, as Rothman, *et al.* suggest, utilize "informal exchanges," "conversations" and other unpublished material to provide information for the parties' contentions in a legal proceeding. Such information, untested by the rigors of the public forum provided by publication, may sometimes be essentially hearsay and gossip, rather than sound science.

Secondly, Rothman, *et al.* seem to forget that the members of a jury are not "practicing scientists". They are seldom familiar with the intricacies of the scientific process; they are unaware of the substrata or characteristics of scientific disputes, and, understandably, they have great difficulty comprehending many of the complexities. Thus, Rothman, *et al.* can easily distinguish the differences between P values and confidence intervals, between substantive (or "quantitative") significance and "statistical" significance, between scientific inference and statistical inference, between a random sample and a randomized trial. Most scientists and almost all physicians, however, have great difficulty in grasping these distinctions. How then is a jury of "laymen" supposed to grasp and apply these concepts unless the judge acts as an active guide, translator, "editor", or referee rather than as passive conduit? I understand that a principal test for the admissibility of expert evidence is that it "assist the trier of fact to understand the evidence or to determine a fact in issue . . ." F.R.E. 702. The admission of "all the facts," as Rothman, *et al.* suggest, untempered by sound guiding principles, would contribute only to confusion or distortion, not understanding.

Rothman, *et al.* have transposed, to a completely inappropriate arena, their admirable thought that scientific information of all kinds, particularly dissenting opinions, should be available for evaluation. The courtroom is not the forum for the advancement of scientific knowledge. It is where juries of the participants' peers assign responsibility for the causation of injuries, not where practicing scientists seek new discoveries.

Finally, Rothman, *et al.* set no standards, give no guidelines, and present no criteria for evaluating the "sources (they) consult as practicing scientists". With no limitation on the kinds of sources that are acceptable, the judge is apparently supposed to allow anything to be examined, including the unpublished re-analysis or meta-

analysis from a "scientific expert" who is no more than a "professional witness".

Because Rothman, *et al.* are presumably thoughtful scientists, I doubt that they have thoroughly considered the consequences of what they propose. Discontent with the existing scientific standards, they apparently seek to regress to an era when no standards at all existed. They suggest that their position is supported by "a large community of respected epidemiologists." I would be surprised if that statement were true. Most members of the "community" of scientists would reject the intellectual anarchy that is being proposed.

The Court's decision in this case will not "alter the role of epidemiological concepts in science" as Rothman, *et al.* suggest. Scientists, not courts, will ultimately determine that role. However, the Court's decision will determine whether scientific standards, designed to ensure the reliability and acceptability of scientific results, can be ignored by scientific witnesses in presenting their testimony to lay juries.

Like editors and reviewers who are the judges in science, legal judges are imperfect. I am certain that they sometimes err or misinterpret. Their rulings are sometimes reversed on appeal, just as the original verdicts of editors and reviewers at one professional journal may be reversed by those at another. Nevertheless, despite their imperfections, judges are the best thing society has available to set and maintain reasonable standards for what is admissible scientific evidence in a court of law.

I do not believe that society, science, or law will be served if judges are rendered unable to carry out that crucial function.

II. PEER REVIEW AND PUBLICATION ESTABLISH NECESSARY STANDARDS AND RIGOROUS DISCIPLINE FOR THE SCIENTIFIC COMMUNITY.

As someone who has regularly published dissents from the "consensus" and the "accepted wisdom," I completely agree with the eloquent assertion by Rothman, *et al.* that "close and repeated questioning of all available scientific information . . . is the engine of scientific progress and the determinant of . . . scientific truths.' Rothman Brief at 12. I also agree with their remark that "peer-reviewed publications have never functioned as . . . sole guarantors of scientific truth." *Id.*

On the other hand, I deplore their failure to offer any better substitute for scientific acceptability than "informal exchange" by "practicing scientists." As an editor, investigator, and author, I know that the peer-review process has many flaws. So does a democratic system of government. But like democracy, peer-review is the best system we have available. It offers the hope, and often the guarantee, that a minimum standard will be maintained for scientific (and literary) discipline.

I doubt that Kenneth Rothman, as editor of a professional journal, automatically accepts and publishes whatever is sent to him; and I know of no other editor (including myself) who would do so. Editors and reviewers are occasionally wrong in their decisions. Even the most prestigious and respected scientific journals have rejected work that later won a Nobel prize, and have accepted work that turned out to be fraudulent, erroneous, or misleading.

Nevertheless, serious scientists know that their work does not become scientific information until it is published. If they believe the work is worthwhile but unappreciated by a set of reviewers at a particular journal, the authors will try elsewhere (perhaps after making some improvements or other changes suggested by the reviewers). Thus, if the work is rejected by Journal A,

the authors try Journal B. If Journal B says "no," they try Journal C. They may go on to many more journals, but eventually they get the work published. If the work is worthwhile, somewhere in the scientific "establishment", the authors will find an editor and/or reviewer who will appreciate—or at least tolerate—it.

All of the famous scientific "heretics" and "nonconformists," from Galileo onward, who had initial difficulty getting their work accepted by the "establishment" that was the "peer-review process" of their era, nevertheless persisted and managed to get the work published. Without that persistence *and publication*, we would not know of the work today.

This is the reason why responsible scientists will almost always decline to rely upon work that has not been published. The absence of publication means that the author has not had the intellectual discipline and self-respect, and enough "confidence" in the work itself, to get it published. Of course, a great deal of valuable information can be involved in the "informal exchanges" of meetings, conversations, letters, or even gossip among scientists. If the information is serious and valuable, however, it becomes expressed more formally in the true "currency" of scientific exchange: publication. An author or investigator who is unable or unwilling to get his work accepted in the fundamental primary "court" of scientific publication should surely not expect to get it accepted secondarily in a court of law.

III. RE-ANALYSIS IS AN ACCEPTABLE SCIENTIFIC METHOD, BUT IT, TOO, SHOULD BE SUBJECTED TO ESTABLISHED SCIENTIFIC STANDARDS AND THE RIGORS OF PEER REVIEW.

The argument that Rothman, *et al.* offer about "re-analysis" is confusing because it encompasses two different activities: (1) re-analysis, discussed in this section, is what occurs when the data or results *for a single study*

are re-appraised; (2) meta-analysis, discussed in the next section, is what occurs when results for a series of studies are "pooled," with conclusions being drawn from the combined aggregate of data.

An effective re-analysis can be done if the analyst has complete access to all the original data of the individual study. For example, the Food and Drug Administration regularly re-analyzes (usually during supervisory checking) the huge collection of original data submitted by pharmaceutical companies as part of new drug applications. In most instances, however, the person doing a re-analysis works not with the original data, but only with the excerpts and summaries that have been published in the scientific literature.

From the available published information, the re-analyst can explore things in diverse ways: putting the cart before the horse, comparing the size of the cart wheels and the number of horses, speculating about what would happen if the cart were larger, had more wheels, or more horses, and so on. Nevertheless, the re-analysis is almost always limited to what was available in the original published report.

This constraint does not prevent re-analysis from being done and from sometimes being quite useful, provided that the analyst has been suitably careful to avoid distortion, and to maintain acceptable scientific discipline in obtaining, presenting, and publishing the results. For example, some years ago, several colleagues and I published, in the *Journal of the American Medical Association*,[3] a re-analysis that demonstrated the fallacies of a study that had originally been published in the *New England Journal of Medicine*. After this re-analysis, and after the appearance of subsequent research that contra-

[3] A. Feinstein, R. Horwitz, W. Spitzer, R. Battista, "Coffee and Pancreatic Cancer: The Problems of Etiologic Science and Epidemiologic Case-Control Research," 246 *Journal of the American Medical Association* 957 (1981).

dicted the original study, and after repeating his own study, the original author eventually "admitted" the error.

I therefore agree with Rothman, *et al.* that re-analysis can furnish acceptable scientific evidence. I disagree, however, with their implicit proposal that judicial acceptance be given to any re-analysis, done in any way the analyst wishes, and never reported in a disciplined scientific manner. If the work is worthy of attention as scientific evidence, it should be submitted to the scientific community for peer review and publication.

IV. META-ANALYSIS PRESENTS SPECIAL PROBLEMS IN ITS AGGREGATION OF DATA; IT IS VITAL THAT STANDARDS BE ESTABLISHED AND MET BEFORE SUCH DATA ARE PRESENTED TO A JURY.

The main statistical purpose of meta-analysis is to obtain statistical significance (or persuasive confidence intervals) for individual studies that did not originally achieve these mathematical accolades. They become achieved in the meta-analysis because the pooling of results for individual studies substantially increases the sample size of the aggregated results, and because significance testing and confidence intervals are so dependent on sample size.

From a scientific standpoint, however, meta-analysis is a highly controversial procedure because it combines heterogeneous entities into a single mixture. Apples, oranges, grapes and bananas may be pooled into a non-specific *fruit salad.* Fruit salad, garden salad, shrimp salad, and pasta salad may be pooled into a non-specific *salad.* Due to these inherent difficulties, scientific dispute has raged, and continues today, over the topic and methods of meta-analysis. *See, e.g.*, "Whither Meta-Analysis," *The Lancet*, April 18, 1987 at 897-898; K. Wachtler, "Disturbed by Meta-Analysis," Vol. 241, *Science*, September 16, 1988, at 1407-1408.

In general, there has been relatively broad but not complete agreement that meta-analyses are justified if the data come from randomized clinical trials. The randomization process can be relied upon (in general) to remove a biased assignment of treatment; and the trial itself will usually have been done with a careful protocol, planned in advance. There is no agreement, however, that meta-analyses are justified if the data come from epidemiologic or other studies that were done without the scientific safeguards of a randomized trial. It is possible, perhaps, to establish scientific standards for selecting reports, adjusting biases, and taking other precautions that might bring a modicum of scientific credibility to meta-analyses of epidemiologic data.[4]

I shall again agree, therefore, with the suggestion by Rothman, *et al.* that meta-analysis should not be automatically excluded from scientific or legal consideration. I shall once again disagree, however, with the laxity of their *laissez faire* permissiveness that would accept any meta-analysis, apparently done by anyone, without any specified standards for the procedure, using whatever "evidence"—published or unpublished—the analyst wants to include, and without any requirement that the analyst present the results to a scientific "jury" (through the process of publication) before they are unveiled in a courtroom. Meta-analysis, probably yes; without any specified scientific standards, surely no.

---

[4] *See e.g.*, Z. Gerbarg and R. Horwitz, "Resolving Conflicting Clinical Trials: Guidelines for Meta-Analysis," 41 *Journal of Clinical Epidemiology* 503 (1987); H. Sacks, J. Berrier, D. Reitman, V. Ancona-Berk, and T. Chalmers, "Meta-Analyses of Randomized Controlled Trials," 316 *New England Journal of Medicine* 450 (1987).

V. CONFIDENCE INTERVALS, LIKE STATISTICAL SIGNIFICANCE, ARE A WAY OF MEASURING NUMERICAL STABILITY. EITHER TOOL MUST BE USED WITH ACCEPTED STANDARDS FOR RESULTS TO BE TRUSTWORTHY FOR EITHER SCIENTISTS OR JURIES.

In their principal argument, Rothman, *et al.* attack reliance upon statistical significance as being too arbitrary, and suggest the use of "confidence intervals" as an alternative. What is important to understand, however, is that both significance testing and confidence intervals are tools for evaluating the stability of the numbers being examined. In urging that significance testing be scientifically expunged and "removed from the inappropriate legal pedestal," Rothman, *et al.* completely ignore the necessity that the data used for drawing inferences be numerically stable, and that specific statistical standards must be established to demonstrate that stability. I shall offer some examples to illustrate the hazards posed by the abandonment of these standards.

A. Numerical Stability.

The numerical stability of quantitative information has been a prominent problem in medical science for at least 40 years. An investigator might claim that a new treatment, having a success rate of 75%, was substantially better than an older treatment, which had a success rate of 50%. Although the difference might initially seem impressive, the reader's enthusiasm might vanish with the discovery that the 75% rate represented only three successes in four patients, whereas the 50% success rate represented two successes in four patients.

These results are numerically unstable. If the next patient in each group fails with the new treatment, and succeeds with the old, the results in the two groups would be identical at 3/5 = 60%. On the other hand, if the original rates of 75% vs. 50% came from groups that each contained 400 patients, so that the comparison was

based on 300/400 vs. 200/400, the numerical results would be stable because the success rate of 75% vs. 50% would hardly be altered by whatever happened to the next patient in each group.

In the examples just cited, the instability of 3/4 vs. 2/4 and the stability of 300/400 vs. 200/400 are evident from visual inspection and a moment's thought. In many other instances, however, the conclusions are not so evident. For the success rates of 75% vs. 50%, stability would be much more difficult to evaluate if the numbers were 12/16 vs. 9/18, or 3/4 vs. 25/50.

B. Null Hypothesis and P Values.

To deal with these issues in numerical stability, scientists began using a statistical method that converted the challenge to an evaluation of probability. The evaluation proces begins with an assumption, called the *null hypothesis*, which postulates that the two treatments (or some other distinguishing feature of the two groups) are essentially identical. If the treatments are identical, and, if there is an observed difference of 75% vs. 50%, then that difference must have occurred by chance.

A mathematical mechanism, called a *P value*, was established to indicate the probability that the observed difference occurred by chance. If the chance probability seemed sufficiently remote, *i.e.*, P was small enough, the assumption that the two treatments were essentially identical would be regarded as untenable. Accordingly, the null hypothesis (with which the statistical reasoning began) would be rejected, and the results would be proclaimed "statistically significant". The implication of this conclusion was that the data were numerically stable, and that the observed difference was "real", *i.e.*, not due to chance.

C. Choice of Alpha Boundary.

To allow standard decisions during this analysis, a boundary level called α (alpha) was needed to separate

the sufficiently small probabilities from those that were "too big". According to recommendations made in 1925 by Sir Ronald Fisher, a giant in the pantheon of statistical methods, α was set at .05. R. Fisher, *Statistical Methods For Research Workers* (Edinburgh; Oliver & Boyd, 1925). After World War II, when medical journals and other evaluators of quantitative data began to concentrate on problems of numerical stability, they decided that any claims of important distinctions must be accompanied by "tests of statistical significance", and they accepted Fisher's boundary level of $\alpha = .05$. This decision was followed by the proliferation of statements in medical literature that "$P \leq .05$", *i.e.*, that in the published numerical results, if the two factors being compared were essentially identical, the probability that the observed distinction could have occurred merely by chance was less than 5%.

The quest for a "good" P value in research soon became so dominant that many readers and evaluators began to ignore the magnitude of the observed results themselves. Instead, if the P value showed "statistical significance", many readers inferred that the results were important and impressive. This inference was often wrong for two reasons: (1) the P value depended exclusively on the numerical data in the compared groups and had nothing to do with the scientific quality of the research design or the acquired evidence; and (2) the P value was immensely affected by the size of the groups (or "samples") under study.

To illustrate the first point, consider two baseball players with batting averages of .350 and .275, respectively. If each player has batted 500 times, so that the difference in batting averages is numerically stable, i.e., statistically significant, one might conclude that the first player was a better hitter than the second. This conclusion would be substantially undermined, however, by information that the second player played for the New York Yankees and the first player played for the Yankees of the Bridgeport,

Connecticut Little League. Statistical significance, standing alone, would be an inadequate basis to draw any scientifically appropriate inferences for the comparison of which player was the better hitter.

D. Role of Sample Size.

Similarly, a large enough sample is no guarantor of importance. If the sample is large enough, any difference, no matter how tiny or trivial, can become "statistically significant."[5] Continuing the baseball analogies, if two baseball players (in the same league) have lifetime batting averages of .283 vs. .284, the difference of .001 is essentially trivial in assessing their relative worth as hitters. Nevertheless, if each player has been to bat about 10,000 times, the difference will be "statistically significant" at $P < .05$.[6]

---

[5] Another demonstration of this point can be made by briefly discussing how a P value is calculated. The discussion is important because the same calculation is used to construct a "confidence interval".

In the first step, a critical ratio, called Z, is formed by dividing the observed difference (such as 75%−50%=25%) by its "standard error".
Thus,

$$Z = \frac{\text{observed difference}}{\text{standard error}} \qquad [A]$$

The standard error is formed by dividing an entity called *standard deviation*, which represents variation in the data, by the square root of a factor that represents the *sample sizes* of the observed groups. Since standard error=(standard deviation)/(sample size factor), the formula for Z can be written as

$$Z = \frac{(\text{observed difference}) \times (\text{sample size factor})}{(\text{standard deviation})} \qquad [B]$$

After a level of α is selected, values of Z_α that correspond to α could be determined (from standard mathematical procedures). If the calculated value of Z (for the observed results) exceeds the chosen value of Z_α, the value of P would be $\leq \alpha$, and would be "statistically significant." Thus, the bigger the value of Z, the smaller would be the value of P, and, in symbols, if $Z \geq Z_\alpha$, then $P \leq \alpha$.

[6] The importance of sample size is evident on inspection of the foregoing formula [B] for Z. Since a big value is desired for Z to

When appraisers of statistical results began to focus on P values (such as $\leq .05$) rather than magnitude of the observed differences (such as .001), the role of science became neglected. Differences of important magnitude might be rejected as "non-significant" because the P value had the disastrous disadvantage of .051, whereas differences of unimportant magnitude might be accepted as "significant" because the P value had the impeccable virtue of .049. Consequently, many scientific writers (including myself) began publishing complaints about the rigidity of an α level of .05, and its concomitant distraction from concentration on the scientific quality of the research.

The complaints did not urge that standards of numerical stability be abandoned. Almost all thoughtful scientists would agree that such standards are necessary, and that $\alpha = .05$ is a reasonable general standard. The argument was mainly that the α level should be somewhat flexible (for situations beyond the scope of the discourse here) and that mathematical P values should not replace the evaluation of the *scientific* quality of the evidence. In their battle against significance testing, Rothman, *et al.* do not at all address the important issues of either numerical stability or scientific quality.

E. Calculation of Confidence Intervals.

Rothman, *et al.* urge a judicial preference for confidence intervals over P values but do not explain why such a preference is needed or justified.

A confidence interval is a way of examining the possible range of a relative risk or other numerical distinction between two studied groups. The interval offers an estimate of the range for the "true" value of the

---

exceed Z_α, and since the sample size factor multiplies the quotient of (observed difference)/(standard deviation), a large enough sample size factor will convert any observed difference into a product large enough to exceed Z_α.

distinction. The interval cannot be constructed without choosing a level for α, and the interval is demarcated with $1 - \alpha$. Thus, if $\alpha = .05$, $1 - \alpha = .95$ and the interval is called a 95% confidence level interval. With a 95% confidence level interval, there is a chance of only 5% that the possible actual distinction between the two studied groups falls outside the range.

The confidence level is an inverse counterpart of decisions with a P value, because both depend on the level chosen for α. The choice of α determines the value of $1 - \alpha$ that is associated with the confidence level. Thus, if $\alpha = .05$, then $1 - \alpha = .95$, and the designation is 95% confidence level. If $\alpha = .1$ or $\alpha = .01$, then the respective designations are 90% confidence interval or 99% confidence interval. Consequently, confidence intervals are an alternative mathematical approach to the probabilistic evaluation of numerical stability. The mathematical reasoning and data are identical with either approach.[7]

Unlike the solitary result of the P value, however, a confidence interval has two values, an upper and lower end of the interval. The lower end produces exactly the same decision as the calculation of a P value. The upper value, however, provides additional information, suggesting how much larger the true difference might have been than what was observed. For this reason, the confidence interval is particularly valuable when an investigator,

---

[7] In the foregoing formula [A] for the calculation of Z, the requirement for statistical significance is that

$$\text{(observed difference)/(standard error)} \geq Z_\alpha \qquad [C]$$

Formula [C] can be rearranged algebraically to show that the requirement for statistical significance is

$$\text{(observed difference)} - (Z_\alpha)\ \text{(standard error)} \geq 0 \qquad [D]$$

Expressed in mathematical terms, the "null hypothesis" value of 0, *i.e.*, no difference in the two treatments, should be excluded from the zone formed in Formula [D] when the product, (Z_α) (standard error), is subtracted from the observed difference. This is the same Z_α used earlier for the decision with formulas [A] or [B].

having found a small observed difference, is tempted to conclude that it is insignificant. In this situation, if the confidence interval is too wide, the investigator cannot conclude there is no difference. For such a situation, further investigation is warranted. In the more customary situation, however, where the investigator finds an impressive difference and wants to claim that the difference is statistically significant, the P value and the confidence interval techniques yield the same conclusion. Because of the added bonus offered by confidence intervals, many investigators (and evaluators) have begun to prefer confidence intervals over P values, although the primacy of the P value is still supported by many prominent statisticians.

The use of confidence intervals, however, does not in any way eliminate the necessity for numerical standards. The most common and widely accepted standard is the use of a 95% confidence interval, which is precisely analogous to a P value of .05, which denotes "statistical significance." The critical issue is what level of α is to be acceptable for scientists in the scientific arena, and for jurors in the legal arena. The choice of α must be made whether statistical significance or confidence intervals are used.

In their argument that confidence intervals should be judicially preferred over the P values of "significance testing," Rothman, *et al.* do not refer to the importance of adherence to accepted standards in selecting the boundary level to measure the stability of numbers. In fact, Rothman, *et al.* mention no standards, no boundaries, and nothing about the importance of a level for α, although such a level must be chosen to allow a confidence interval to be calculated. With no constraints on levels of α, data analysts would apparently be free to choose these levels in any way they want. The choice of α (or $1-\alpha$) can make the confidence intervals as large or as small as the investigator wishes, and can thus make the data support whatever results the data analyst would like to demonstrate.

Thus, Rothman, *et al.* are apparently urging that the occasionally excessive rigidity of P values be replaced by the anarchy of unrestricted confidence intervals. Whatever the statistical impact of this change, of course, neither the P value nor the confidence interval addresses the fundamental scientific necessity of quality in the research design and data.

Rothman, *et al.* suggest that their position on numerical stability is supported by "a large community of respected epidemiologists." Rothman Brief at 8. If so, I suspect that the epidemiologists do not really understand precisely what is implied in Rothman, *et al.*'s recommendations. I feel certain that most members of the "community" of scientists would reject the numerical anarchy that is being proposed. They would insist that standards of intellectual discipline be honored, that standards of numerical stability be maintained, and that evidence not be accepted if it fails to meet the standards of numerical stability. The standards can be established either via P values or circumscribed confidence intervals (or even with some new "fragility" techniques that are beyond the scope of this discussion). The standards should be appropriately flexible, but they must exist if science is to preserve its tradition of intellectual discipline and high quality research.

I can think of no better way to allow "junk science" into both the published literature and the courtroom than to remove the constraints that demand numerical stability for data. If researchers can choose confidence intervals in an *ad hoc* manner, without justifying their selection of appropriate boundary values according to accepted scientific standards, the result will be confidence games, not confidence intervals.

CONCLUSION

Standards and peer review, especially as reflected in published scientific work, are vital to the understanding of practicing scientists in their own work. Such requirements are equally important if scientific information is to be used when juries make decisions. If scientific and statistical standards are abandoned, and if juries receive untrustworthy information, courtroom decisions will be based on unfettered speculation and conjecture.

Respectfully submitted,

DON M. KENNEDY
LORETTA M. SMITH
RICHARD A. OETHEIMER
Counsel of Record
GOODWIN, PROCTER & HOAR
Exchange Place
Boston, MA 02109
(617) 570-1000
Counsel for Amicus Curiae

January 19, 1993

APPENDIX

IDENTIFICATION OF AMICUS CURIAE

Alvan R. Feinstein was born in Philadelphia in 1925. He received his Bachelor of Science degree, a Master's Degree in Mathematics, and his Doctor of Medicine degree from the University of Chicago. After residency training in internal medicine at Yale University and at Columbia-Presbyterian Hospital in New York, and after a research fellowship at the Rockefeller Institute, he became medical director at Irvington House, just outside of New York City, where he studied a large population of patients with rheumatic fever. In this research, he began developing new clinical investigative techniques that were eventually expanded beyond rheumatic fever into many other activities, particularly work on the prognosis and therapy of cancer.

His new clinical epidemiologic approaches and methods were first reported in two books, *Clinical Judgment* (Williams & Wilkins Co., 1967, Reprinted, Robert E. Krieger Publishing Co., 1974) and *Clinical Biostatistics* (C.V. Mosby Co., 1977). His two latest books, *Clinical Epidemiology* (W.B. Saunders Co., 1985) and *Clinimetrics* (Yale University Press, 1987), describe the structure and contents of clinical research with groups, and the strategy used to form clinical indexes and rating scales for important human clinical phenomena—such as pain, distress, and disability—that have not received adequate attention in an age of technologic data.

To supplement the current biomedical forms of "basic science" that are used for explanatory decisions about pathophysiologic mechanisms of disease, Dr. Feinstein has vigorously advocated that clinical epidemiology and clinimetrics be developed as an additional humanistic "basic science" for the managerial decisions of clinical practice.

Dr. Feinstein was instrumental in pioneering the field of clinical epidemiology. He has cautioned against overreliance on statistical results and stressed the need for attention to the descriptive data that are the foundation of clinical therapy and to basic scientific principles in the collection and analysis of data. From 1970 to 1974, Dr. Feinstein served on the Biomedical Epidemiological Methodology Advisory Committee of the Food and Drug Administration. Since that time, Dr. Feinstein has continued to advocate the development of a well delineated set of scientific standards, based on rigorous application of accepted scientific principles, for conducting epidemiologic investigations.

Dr. Feinstein is Sterling Professor of Medicine and Epidemiology at the Yale University School of Medicine, where he is also Director of the Clinical Epidemiology Unit, the Robert Wood Johnson Clinical Scholars Program, and the Introduction to Clinical Medicine Course (for second-year medical students). He has held numerous visiting professorships at universities in this country and abroad, and he served a six-year term (1977-1983) as a member of the governing board of the American Board of Internal Medicine. He has been elected to the American Epidemiological Society and to the Association of American Physicians, being one of a few medical academicians chosen for both of these prestigious groups.

Dr. Feinstein has been awarded many honors for his work in the fields of medicine and public health. These honors include the Richard and Hinda Rosenthal Foundation Award from the American College of Physicians in 1982 for "making a notable contribution to improve clinical care in the field of internal medicine"; the Robert J. Glaser Annual Award of The Society for Research and Education in Primary Care Internal Medicine in 1987 for "exceptional contributions to research and education in the field of general internal medicine"; and the J. Allyn Taylor International Prize in Medicine in 1987, for "sig-

nificant contributions to a field of basic or clinical research" (shared with Peter Armitage and David L. Sackett).

Dr. Feinstein has served on the editorial boards of many medical and scientific journals, including the *Journal of Chronic Diseases*, the *Archives of Internal Medicine*, the *Journal of the American Medical Association*, *The Statistician*, and *Clinical Pharmacology and Therapeutics*. He is past editor of the *Journal of Chronic Diseases* and is currently co-editor of the *Journal of Clinical Epidemiology*.

Dr. Feinstein has written or co-authored more than 300 articles that have appeared in the medical and scientific literature during the last 40 years.